W460–W464 *Nucleic Acids Research, 2005, Vol. 33, Web Server issue*
doi:10.1093/nar/gki456

BABELOMICS: a suite of web tools for functional annotation and analysis of groups of genes in high-throughput experiments

Fátima Al-Shahrour[1], Pablo Minguez[1], Juan M. Vaquerizas[1], Lucía Conde[1] and Joaquín Dopazo[1,2,*]

[1]Bioinformatics Unit, Centro Nacional de Investigaciones Oncológicas (CNIO), Melchor Fernández Almagro 3, 28029, Madrid, Spain and [2]Functional Genomics Node, INB, Centro de Investigación Príncipe Felipe, Autopista del Saler 16, E46013, Valencia, Spain

Received February 14, 2005; Revised and Accepted April 4, 2005

Downloaded from http://nar.oxfordjournals.org/ at Michigan State University on November 25, 2015

ABSTRACT

We present Babelomics, a complete suite of web tools for the functional analysis of groups of genes in high-throughput experiments, which includes the use of information on Gene Ontology terms, interpro motifs, KEGG pathways, Swiss-Prot keywords, analysis of predicted transcription factor binding sites, chromosomal positions and presence in tissues with determined histological characteristics, through five integrated modules: FatiGO (fast assignment and transference of information), FatiWise, transcription factor association test, GenomeGO and tissues mining tool, respectively. Additionally, another module, FatiScan, provides a new procedure that integrates biological information in combination with experimental results in order to find groups of genes with modest but coordinate significant differential behaviour. FatiScan is highly sensitive and is capable of finding significant asymmetries in the distribution of genes of common function across a list of ordered genes even if these asymmetries were not extreme. The strong multiple-testing nature of the contrasts made by the tools is taken into account. All the tools are integrated in the gene expression analysis package GEPAS. Babelomics is the natural evolution of our tool FatiGO (which analysed almost 22 000 experiments during the last year) to include more sources on information and new modes of using it. Babelomics can be found at http://www.babelomics.org.

INTRODUCTION

Molecular biology has typically addressed functional questions by studying individual genes, either independently, or a few at a time. Despite the intrinsic reductionism of this approach, an important part of our knowledge on functional properties and biological roles of genes and gene products was obtained in this way. Nevertheless, the possibility of obtaining information on thousands of genes or proteins using high-throughput methodologies, such as gene expression (1) and proteomics (2), has opened up new avenues in querying living systems at the genome level that are beyond the old paradigm 'one-gene-one-postdoc'. Relevant biological questions regarding gene, gene product interactions or biological processes played by networks of components, etc. can now, for the first time, be addressed realistically. Nevertheless, caution must be exercised when dealing with these excess data because spurious associations may arise if proper methodologies are not used [for discussions in some related aspects see (3,4)].

Unfortunately, these spurious associations are often considered as evidence of actual functional links, leading to misinterpretation of results. All these features of genomic data must be taken into account for any procedure aiming at properly identifying functional roles in groups of genes. Pursuing this goal, several years ago, we developed the FatiGO (Fast Assignment and Transference of Information using Gene Ontology (GO), available at http://www.fatigo.org) tool (5). FatiGO was the pioneering tool for finding significant differences in the distribution of GO terms between groups of genes taking the multiple testing nature of the contrast into account to avoid the above mentioned spurious associations. One of the main fields of application of FatiGO has been the analysis of gene expression microarray data. A clear example is the study of gene co-expression, which tends to be an evidence of

*To whom correspondence should be addressed. Tel: +34 96 3289680; Fax: +34 96 3289701; Email: jdopazo@cnio.es
Present address:
Fátima Al-Shahrour, Pablo Minguez, Juan M. Vaquerizas, Lucía Conde and Joaquín Dopazo, Bioinformatics Department, Centro de Investigación Príncipe Felipe, Autopista del Saler 16, E46013, Valencia, Spain

© The Author 2005. Published by Oxford University Press. All rights reserved.

The online version of this article has been published under an open access model. Users are entitled to use, reproduce, disseminate, or display the open access version of this article for non-commercial purposes provided that: the original authorship is properly and fully attributed; the Journal and Oxford University Press are attributed as the original place of publication with the correct citation details given; if an article is subsequently reproduced or disseminated not in its entirety but only in part or as a derivative work this must be clearly indicated. For commercial re-use, please contact journals.permissions@oupjournals.org

common function (6). The use of tools such as FatiGO is becoming essential for the interpretation of microarray experiments and can be applied to any other type of experiment involving a large number of genes (proteomics, interatomics, etc.) (7).

Extending both the sources and uses of the information has helped us develop Babelomics, a suite of web tools for functional annotation and analysis of groups of genes in high-throughput experiments. Currently, Babelomics is composed of six modules and includes biological information for functional annotation coming from different sources, such as GO (8), pathways (9), Interpro functional motifs (10), tissues and chromosomal locations.

Babelomics can be found at http://www.babelomics.org.

THE BABELOMICS RESOURCE

The resource presented in this paper is named after the tale, 'The Babel library' (11), by the famous Argentinean writer Jorge Luís Borges. In the tale, an infinite library is described: 'The universe (which is referred to as the Library) is composed of an indefinite and perhaps infinite number of hexagonal galleries … There are five shelves for each of the hexagon's walls; each shelf contains thirty-five books of uniform format; each book is four hundred and ten pages; each page of forty lines, each line, of some eighty letters, which are black in color'. Such an infinite library would have any book, but would also have an infinite combination of meaningless letters. Finding the real books among the pile of meaningless texts is an excellent metaphor for the challenge that constitutes the extraction of information out of the mass of data in the post-genomic era. What is real and what is just an association by chance is an important issue when dealing with 'omics' data (3,4). Since there is a possibility of high occurrence of spurious associations if proper methodologies are not used, these abundant data must be carefully considered when trying to assign functional properties to groups of genes (7). Babelomics offers different procedures to significantly associate distinct functional labels to groups of genes within the proper statistical frame.

FatiGO/FatiWise IN THE CORE OF BABELOMICS

Testing unequal distribution of terms between two groups of genes

FatiGO (5) takes two lists of genes (ideally a group of interest and the rest of the genome, although any two groups, formed in any way, can be tested against each other) and convert them into two lists of GO terms using the corresponding gene-GO association table. A Fisher's exact test for 2 × 2 contingency tables is used. For each GO term, the data are represented as a 2 × 2 contingency table with rows being presence/absence of the GO term, and each column representing each of the two clusters (so that the numbers in each cell would be the number of genes of the first cluster where the GO term is present, the number of genes in the first cluster where the GO term is absent, and so on). In addition to GO, any other functional label can be used. For example, FatiWise implements InterPro motifs (10), KEGG (9) pathways and Swiss-Prot keywords.

The structure of the functional labels has an important impact on the strategy for performing the test. For example, InterPro motifs have a 'flat' organization with a correspondence of one or more motifs per protein. Terms in GO, on the other hand, have a hierarchical structure called DAG (for directed acyclic graph, where each term can have one or more child terms as well as one or more parent terms). Terms at higher levels, close to the root, of the hierarchy describe more general functions or processes while terms at lower levels are more specific. The level at which a gene is annotated in the GO hierarchy depends on the knowledge of the details of its biological behaviour the annotator had. Testing terms organized in such a way poses an additional difficulty because in some cases they are not exclusive but only constitute descriptions of the same behaviour at different levels of detail (e.g. where is the point in testing apoptosis versus regulation of apoptosis?). To deal with this, FatiGO implements an inclusive analysis (5), in which a level in the DAG hierarchy is chosen for the analysis. Genes annotated with terms that are descendant of the term corresponding to the level chosen, therefore, take the annotation from the parent. If the level corresponding to, for example, apoptosis was selected, any gene annotated as either apoptosis or as any children term was considered in the same category (apoptosis) for the test. This increases the power of the test. There are less terms, each with more genes, to be tested (5,7).

What can be considered a significant biological difference? Statistical approaches and the multiple testing problem in Babelomics

As mentioned previously, great caution should be exercised when dealing with a large set of data because of the high occurrence of spurious associations (3).

Addressing multiple testing properly is a rather complex problem. Many of the conventional correction methods (e.g. Bonferroni or Sidak) are based on the consideration that a P-value should be adjusted by multiplying a reasonable significant threshold (e.g. $P < 0.05$) for the number of tests performed to obtain a new threshold. When thousands of tests are performed the original assumption risks to be too conservative.

A better strategy to estimate P-values is provided by another family of methods allowing less conservative adjustments, such as the family wise error rate (FWER); this controls the probability that one or more of the rejected hypotheses (GO terms whose differences cannot be attributed to chance) is true (i.e. a false positive). The minP step-down method (12), a permutation-based algorithm, provides a strong control (i.e. under any mixture of false and true null hypothesis) of the FWER. Approaches that control the FWER can be used in this context although they are dependent on the number of hypotheses tested and tend to be too conservative for a high number of simultaneous tests. Aside from a few cases in which FWER control could be necessary, the multiple testing problem in functional assignation does not require protection against even a single false positive. In this case, the drastic loss of power involved in such protection is not justified. It would be more appropriate to control the proportion of errors among the identified GO terms, whose differences among groups of genes cannot be attributed to chance instead. The expectation of

Downloaded from http://nar.oxfordjournals.org/ at Michigan State University on November 25, 2015

this proportion is the false discovery rate (FDR). Different procedures offer strong control of the FDR under independence and some specific types of positive dependence of the tests statistics (13), or under arbitrary dependency of test statistics (14).

All the modules in Babelomics return adjusted *P*-values based on three different methods of accounting for multiple testing as explained above (12–14).

TransFAT: the extension of FatiGO/FatiWise to the study of transcription factor binding sites

TransFAT (Transcription Factor Association Test) is a tool designed to detect under or over-representation of putative transcription factors binding sites (TFBSs) in sets of genes, by comparing them against a set of reference (e.g. a cluster of co-expressing genes against the rest of the genes in the analysis). Particular transcription factors (TFs) are assigned to genes if the corresponding predicted TFBSs for that TF if found in the 10 kb 5′ region of the gene. Search is carried out by the Match program (15), using only high-quality matrices and with a cut-off to minimize false positives, from the Transfac database (16).

If the experiment implies finding the TF (among the many possible) involved in the activation, then a final list of significantly over-represented TFBSs is provided with adjusted *P*-values associated as explained above. Nevertheless, it is quite common that some information is available beforehand on the possible TF involved in the regulation of the genes. In this case, if only one TF is checked, the *P*-value does not need any adjustment. The present version only includes predictions of TFBSs for human sequences, but future versions will include more organisms.

Tissues Mining Tool (TMT): introducing another dimension in the functional assignments

TMT is a web application to extract significant information related to the differential expression of two sets of genes in tissues.

Gene expression data are taken from the SAGE Tag libraries downloaded from the Cancer Genome Anatomy Project (CGAP, http://cgap.nci.nih.gov/). A total of 279 human libraries that belong to 29 different tissues and 190 mouse libraries from 26 tissues are used.

TMT compares frequency gene expression of two lists of genes in the tissues selected using a *t*-test. Users may submit two lists of genes in unigene cluster annotation, select a set of tissues and histology categories and a *t*-test will be performed by TMT for each of the libraries included in the selection. Options to filter quality of libraries are provided. The tool obtains a frequency matrix that can also be filtered selecting maximum percentage for null values accepted in rows (libraries) and columns (genes).

The TMT's second application is to compare gene expression significance between a user submitted list and the genes with frequency expression data in the tissue and histology category selected. This option permits the comparison of lists of genes obtained in any experiment to standard representation of gene expression in tissues in a given histological category.

The output of the program shows a number of genes with non-null frequency values in libraries per list, *t*-test results, a color-coded image containing a visual representation of the above result and a set of informative tables.

GenomeGO: mapping function onto the genome

GenomeGO is a web interface for the inference of over or under-representation of GO terms in the same way as FatiGO, but in this case the query cluster contains the genes present in a chromosomal region and reference cluster is the total of genes of the genome (not including those in the query cluster). GenomeGO implements Fisher's exact test for 2 × 2 contingency tables for comparing two groups of genes (query and the rest of genes of the genome). Similar to FatiGO, the result is a list of GO terms with a significant different distribution among the groups. The results of the test are corrected for multiple-testing using the methods described above to obtain adjusted *P*-values.

The genes of the query group can be selected directly from their location in a region of the genome, or can be just provided as a list. Genomic regions can be selected either by defining a range of chromosome coordinates or by directly choosing the cytoband of interest. Gener corresponding to the selected chromosomal regions age was located using the Ensembl (17).

GenomeGO is very useful to study chromosomal regions at the functional level. If some alteration is found in a chromosome region, the functions of the genes therein can be studied through this tool. GenomeGO is the functional complement to other tools for exploring chromosomal alterations, such as InSilicoCGH (18).

FatScan: studying correspondences between phenotypes and molecular roles of genes by analysing ordered lists of genes

The analysis of genome-scale data from different high-throughput techniques can be used to obtain lists of genes ordered according to their different behaviours under distinct experimental conditions corresponding to different phenotypes (e.g. differential gene expression between diseased samples and controls and different response to a drug). The order in which the genes appear in the list is a consequence of the biological roles the genes are carrying out within the cell, accounting, at molecular scale, for the macroscopic differences observed between the phenotypes studied. Typically, two steps are followed for understanding the biological processes that differentiate phenotypes: first, genes with significant differential expression are selected on the basis of their experimental values and subsequently the functional properties of these genes are analysed by using, e.g. any of the procedures described above. This procedure is applied when, for example, genes differentially expressed are sent from the pomelo tool (18,19) to the FatiGO in order to check for the biological roles carried out by these genes.

Instead, we present a simple procedure which combines experimental measurements with available biological information in a way that genes are simultaneously tested in groups related by common functional properties. It constitutes a very sensitive tool for selecting genes with significant differential behaviour in the experimental conditions tested.

The procedure consists in the application of a test for extracting significant over- or under-represented terms associated with groups of genes (5). The test is sequentially applied to

Downloaded from http://nar.oxfordjournals.org/ at Michigan State University on November 25, 2015

different partitions of an ordered list of genes previously obtained by studying their differential expression according to the phenotypes analysed. Then, labels for biological information are used to test whether groups of genes sharing these labels are simultaneously over or under expressed or, in other words, if these genes tend to be concentrated at one end of the list instead of being uniformly distributed across it. Contrary to GSEA (20), and other alternative methods based on comparison of distributions through Kolmogorov–Smirnov or related tests, this method does not require an extreme non-uniform distribution of genes. It is able to find different types of asymmetries in the distributed groups of genes across the list of data taking into account the strong multiple-testing nature of the contrast.

We propose the use of such a procedure to scan ordered lists of genes for the understanding of the biological processes operating behind them. This procedure can be useful in situations in which it is not possible to obtain statistically significant differences based on the experimental measurements (low prevalence diseases, etc.).

DISCUSSION

The FatiGO (available at http://www.fatigo.org) tool was the first application for finding significant differences in the distribution of GO terms between groups of genes taking into account the multiple testing nature of the contrast (5). During the last year, FatiGO has analysed almost 22 000 experiments and has reached an average of 65 daily users. The average use per countries is: US (domains edu 15% and net 8%) 23%, France 9%, Spain 6%, UK 5%, etc. These ratios of use, as well as the user's profile show the interest for functional annotation tools for groups of genes beyond the traditional one-gene-at-a-time approach. The scope of the information implemented in such statistical framework was initially extended to interpro motifs (10), KEGG pathways (9) and Swiss-Prot keywords in the FatiWise (18). In this new version, we present a complete suite including the analysis of TFBS, chromosomal positions and presence in tissues with determined histological characteristics with the help of three new modules, TransFAT, GenomeGO and TMT, respectively. Additionally, the module FatiScan provides a new procedure which integrates biological information in combination with the experimental results in order to find groups of genes with modest but coordinate significant differential behaviour. FatScan constitutes an alternative to GSEA (20), because it is capable of finding asymmetries in the distribution of genes even if they are not extreme.

The suite of tools for functional annotation of groups of genes, Babelomics, can be used in combination with GEPAS tools (18,19) to study gene expression using microarrays. All the tools in Babelomics are connected to the proper tools in GEPAS. Additionally, the tools can be used to analyse any type of high-throughput or genome-scale experiment in which the interest is in the behaviour of groups of genes.

ACKNOWLEDGEMENTS

L.C. and this work is supported by grant PI020919 from the FIS. F.A.-S. is supported by grant BIO2001-0068 from MCyT. J.M.V. is supported by a FPU fellowship from the MEC. P.M. is supported by a grant from Genoma España and Canada Genome. This work is also partly supported by grants from Fundación Ramón Areces, Fundació La Caixa and Fundacion BBVA and RTICCC from the FIS. Fundación Genoma España supports the Functional Genomics node (INB). IBM awarded us with a SUR grant. Funding to pay the Open Access publication charges for this article was provided by Genoma España.

Conflict of interest statement. None declared.

REFERENCES

1. Holloway,A.J., van Laar,R.K., Tothill,R.W. and Bowtell,D.D. (2002) Options available-from start to finish—for obtaining data from DNA microarrays II. *Nature Genet*., **32** (Suppl.), 481–489.
2. MacBeath,G. (2002) Protein microarrays and proteomics. *Nature Genet*., **32** (Suppl.), 526–532.
3. Ge,H., Walhout,A.J. and Vidal,M. (2003) Integrating 'omic' information: a bridge between genomics and systems biology. *Trends Genet*., **19**, 551–560.
4. Díaz-Uriarte,R. (2005) Supervised methods with genomic data: a review and cuationary view. In Azuaje,F. and Dopazo,J. (eds), *Data Mining and Visualisation Methods for Integrative Biology*. John Wiley and Sons, pp. 193–214.
5. Al-Shahrour,F., Díaz-Uriarte,R. and Dopazo,J. (2004) FatiGO: a web tool for finding significant associations of Gene Ontology terms to groups of genes. *Bioinformatics*, **20**, 578–580.
6. Lee,H.K., Hsu,A.K., Sajdak,J., Qin,J. and Pavlidis,P. (2004) Coexpression analysis of human genes across many microarray data sets. *Genome Res*., **14**, 1085–1094.
7. Al-Shahrour,F. and Dopazo,J. (2005) Ontologies and functional genomics. In Azuaje,F. and Dopazo,J. (eds), *Data Mining and Visualisation Methods for Integrative Biology*. John Wiley and Sons, pp. 99–112.
8. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al*. (2000) Gene Ontology: tool for the unification of biology. *Nature Genet*., **25**, 25–29.
9. Kanehisa,M., Goto,S., Kawashima,S., Okuno,Y. and Hattori,M. (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res*., **32**, D277–D280.
10. Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Barrell,D., Bateman,A., Binns,D., Biswas,M., Bradley,P., Bork,P. *et al*. (2003) The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Res*., **31**, 315–318.
11. Borges,J.L. (2000) The library of Babel. *Inquisitions*. Penguin Books Ltd, London (original: Borges,J.L. (1941) La Biblioteca de Babel, Ficciones. Ed. Mar del Plata).
12. Westfall,P.H. and Young,S.S. (1993) *Resampling-Based Multiple Testing*. John Wiley and Sons, NY.
13. Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B*, **57**, 289–300.
14. Benjamini,Y. and Yekutieli,D. (2001) The control of the false discovery rate in multiple testing under dependency. *Ann. Statist*., **29**, 1165–1188.
15. Kel,A.E., Gößling,E., Reuter,I., Cheremushkin,E., Kel-Margoulis,O.V. and Wingender,E. (2003) MATCHTM: a tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res*., **31**, 3576–3579.
16. Wingender,E., Chen,X., Hehl,R., Karas,H., Liebich,I., Matys,V., Meinhardt,T., Pruss,M., Reuter,I. and Schacherer,F. (2000) TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Res*., **28**, 316–319.
17. Birney,E., Andrews,T.D., Bevan,P., Caccamo,M., Chen,Y., Clarke,L., Coates,G., Cuff,J., Curwen,V., Cutts,T. *et al*. (2004) An overview of Ensembl. *Genome Res*., **14**, 925–928.
18. Herrero,J., Vaquerizas,J.M., Al-Shahrour,F., Conde,L., Mateos,Á., Santoyo,J., Díaz-Uriarte,R. and Dopazo,J. (2004) New challenges in gene

Downloaded from http://nar.oxfordjournals.org/ at Michigan State University on November 25, 2015

expression data analysis and the extended GEPAS. *Nucleic Acids Res.*, **32**, W485–W491.

19. Herrero,J., Al-Shahrour,F., Díaz-Uriarte,R., Mateos,A., Vaquerizas,J.M., Santoyo,J. and Dopazo,J. (2003) GEPAS: a web-based resource for microarray gene expression data analysis. *Nucleic Acids Res.*, **31**, 3461–3467.
20. Mootha,V.K., Lindgren,C.M., Eriksson,K.F., Subramanian,A., Sihag,S., Lehar,J., Puigserver,P., Carlsson,E., Ridderstrale,M., Laurila,E. *et al.* (2003) PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genet.*, **34**, 267–273.

Downloaded from http://nar.oxfordjournals.org/ at Michigan State University on November 25, 2015